Detection of HIV Hybrid Sequences Using VESPA

Andrew Farmer and Gerald Myers

Theoretical Biology and Biophysics,
MS K710, Los Alamos National Laboratory, Los Alamos NM 87545

INTRODUCTION

It is well-known that HIV's extremely plastic genome creates a multitude of problems for analysis of sets of sequence data. The most pervasive of these problems is **homoplasy**—the chance occurrence of identical characters (nucleotides or amino acids) in sequences of otherwise different lineages. Character **weighting** and **pruning** (Wills 1995) are two attempts to reduce the level of homoplasy in data sets, and thereby to increase the signal-to-noise ratio; VESPA (viral epidemiology signature pattern analysis) is yet another. In this section, we explore the use of VESPA to evaluate HIV-1 sequences suspected of being intersubtype, or intrasubtype, mosaics or recombinants. This approach differs in several ways from other approaches described in this compendium, hence it offers a complementary and critical test for mosaicism. Furthermore, VESPA, unlike some of the other computational strategies, exists as a relatively simple program that runs on Unix, Macintosh or PC-Dos machines and is available at no cost.

In any set of aligned homologous sequences, positions at which no variation is present are uninformative for most purposes. To be an **informative site**, two or more distinct characters must be present at a site, and each of these distinct characters must be present in two or more of the sequences (Li and Graur, 1991). In other words, different characters must be distributed among different sequences such that at least two different proposed phylogenies will require a different number of mutational events to explain the character distribution at that position; when this is so, the position is considered to be informative, insofar as it provides a reason for choosing one phylogeny over another according to the principle of parsimony.

Homoplasy is often encountered at informative sites, affecting tree analyses and bootstrapping, therefore pruning strategies are invoked in an attempt to identify subsets of varied sites with the least homoplasy. VESPA is such a strategy in so far as it identifies arrays of **atypical** characters in subsets of homologous (and homoplastic) sequences. This is illustrated through the use of VESPA in the widely-known Florida dentist case (Acer-Bergalis case), which involved HIV-1 *env* V3 region sequences of the dentist, his infected patients, Florida control individuals harboring HIV, and HIV-infected individuals from the U.S. at large (Ou et al., 1992; Korber and Myers, 1992). The dentist's viral signature pattern, a noncontiguous array of atypical characters (nucleotides or amino acids) was determined against the U.S. set of subtype B sequences. A set of Florida control sequences was then scanned using the dentist's viral signature to determine the extent of sharing of these atypical characters: no other Florida sequence, excluding the dentist's patients, possessed as many as seven of the dentist's thirteen nucleotide signature characters; most Florida and U.S. sequences shared only one or two of these characters. On the other hand, the patients thought to have been infected through him carried one or more sequences with all of these characters. Note that of the 300 or so total alignable sites, of which perhaps as many as half were varied sites, merely thirteen characters were identified as **differentiating**, and even these involved a low level of homoplasy.

In the HIV recombination problem, "undirected convergence", or homoplasy (Stewart, 1993) will have a significant impact upon attempts to screen for hybrid sequences. Lack of divergence will also contribute to uncertainty over stretches. For this reason, we have elected to explore VESPA as a **detection** tool primarily; the **scoring** and statistical dimensions of its output will be discussed at a later time.

METHODS

HIV-1 mosaicism can be examined as an intergenic or an intragenic phenomenon, and the focus can be intersubtypic or intrasubtypic. The analyses below are concerned with intragenic *env* mosaicism at both intrasubtypic and intersubtypic levels using amino acid characters only. More than 100 complete, alignable HIV-1 *env* amino acid sequences of length approximately 850 characters are available for analysis in 1995. Although amino acid residues offer fewer sites for analysis than nucleotides, the possibilities for reducing homoplasy are greater with twenty rather than four characters. Gaps are included as signature sites.

In the case of intersubtype analysis, there are two parameters, p and q, which are relevant to the determination of differentiating characters for each subtype (A-G). The first of these specifies the degree of conservation of the majority amino acid in the query subtype, establishing a minimum level for the observed frequency of the signature (atypical) character in the query set. The second parameter, on the other hand, specifies the upper limit on the frequency with which the majority query residue may be observed in the background subtype, thus removing from the signature those positions for which a significant number of sequences in the background subtype have the same residue as the majority of sequences in the query subtype, for whatever reason (homoplasy or simple lack of divergence). For example, to define a signature pattern for subtype A using the values 75 and 20 for p and q respectively, the set of A subtype sequences will be compared to each other subtype in turn, resulting in 6 sets of characters. Each signature character in these sets will satisfy the p criterion by being observed in the subtype A sequences at that position at least 75% of the time. Each set will also be characterized by the fact that the background subtype against which it was determined will display the signature character at the corresponding position in no more than 20% of its sequences. These sets of characters individually distinguish subtype A from each of the other subtypes. By taking a consensus of the six sets, requiring that each character in the consensus signature distinguishes the A subtype from each of the six other subtypes, we obtain the A subtype signature pattern. A necessary and sufficient condition to ensure the uniqueness of signature characters for each of the subtypes using this method is $p \leq q$ (Appendix I). The relative stringency of this overall approach becomes evident when merely 20 or so differentiating characters are defined (say for the 50,10 settings) for a given subtype out of the 850 or so total positions.

Intrasubtype analyses pose a different problem, as we do not have pre-defined clusters of sequences to run against each other. Presumably, one could use an approach similar to the intersubtype method by defining clades within the subtype of interest, but in general, the structure of intrasubtype clades is much less well defined than on the level of the subtypes. An obvious alternative to this approach is to define signature patterns for each sequence in the set against all of the other sequences. Under this method, the parameter p becomes irrelevant, as the query set consists of only one sequence and must *a fortiori* satisfy any possible level of conservation in the query set. On the other hand, it is important to exclude from consideration all positions for which the set of all sequences displays minimal conservation, as these positions will tend to exhibit only noise. Therefore, we introduce a new parameter r, which specifies that the background set must be conserved to a certain minimum degree in order for the position to be considered for the sequence signature pattern. The parameter q is also used, again specifying a maximum frequency that the character in the query sequence may be observed in the background set at the same position. Thus, for example, a signature pattern determined for a sequence belonging to a set of 100 other sequences, using the values 75,20 for r and q respectively, would include only those characters for which at least 75 sequences in the background set were in agreement with one another at that position, and fewer than 20 sequences in the background set shared the character shown by the query sequence at that position. This general approach has been found successful for analyzing papillomavirus sequences (Farmer and Myers, 1995).

There are unavoidable violations of the "independence assumption" in this analysis, as in virtually all sequence analyses: specifically, mosaic sequences may be members of the query over which a signature pattern is determined and they may be members of the background.

The VESPA program is available at no cost from the HIV Sequence Database (kam@t10.lanl.gov or cxc@t10.lanl.gov).

RESULTS

A. Intersubtype Analyses

VESPA was run under various threshold conditions of *p,q* for each of the *env* sequence subtype sets using all other subtype sequences taken sequentially as described above. For any given pair of thresholds, certain subtype sets, such as E and F, produced a high number of signature (differentiating) characters whereas others, such as subtype A, produced a low number. These differences, which could be as great as five-fold, were mostly due to sampling and phylogenetic effects—Es and Fs being less diverse than As. As a result of these discrepancies, when a query sequence, a putative hybrid, was run against the various subtypes, a false impression of "E-ness" could be obtained as a straightforward consequence of homoplasy in conjunction with an unbalanced number of signature characters (Fig. 1a). (This effect, which is preeminently obvious through VESPA analysis, can be a less obvious but equally troublesome effect in other approaches.). By employing a variable set of thresholds, and ensuring that divergent members of each of the subtypes were appropriately included (*e.g.* the CAR E sequences with the Thai E sequences), a balanced set of differentiating characters was attained and false impressions of mosaicism, say with E, were eliminated (Fig. 1b). The particular thresholds used in our analysis were:

subtype	*p*	*q*
A	50	35
B	50	20
C	70	20
D	50	20
E	75	10
F	75	10
G	65	20

In general, the greater the diversity of a subtype sequence set (A), the less stringent the selection in order to attain differentiating characters.

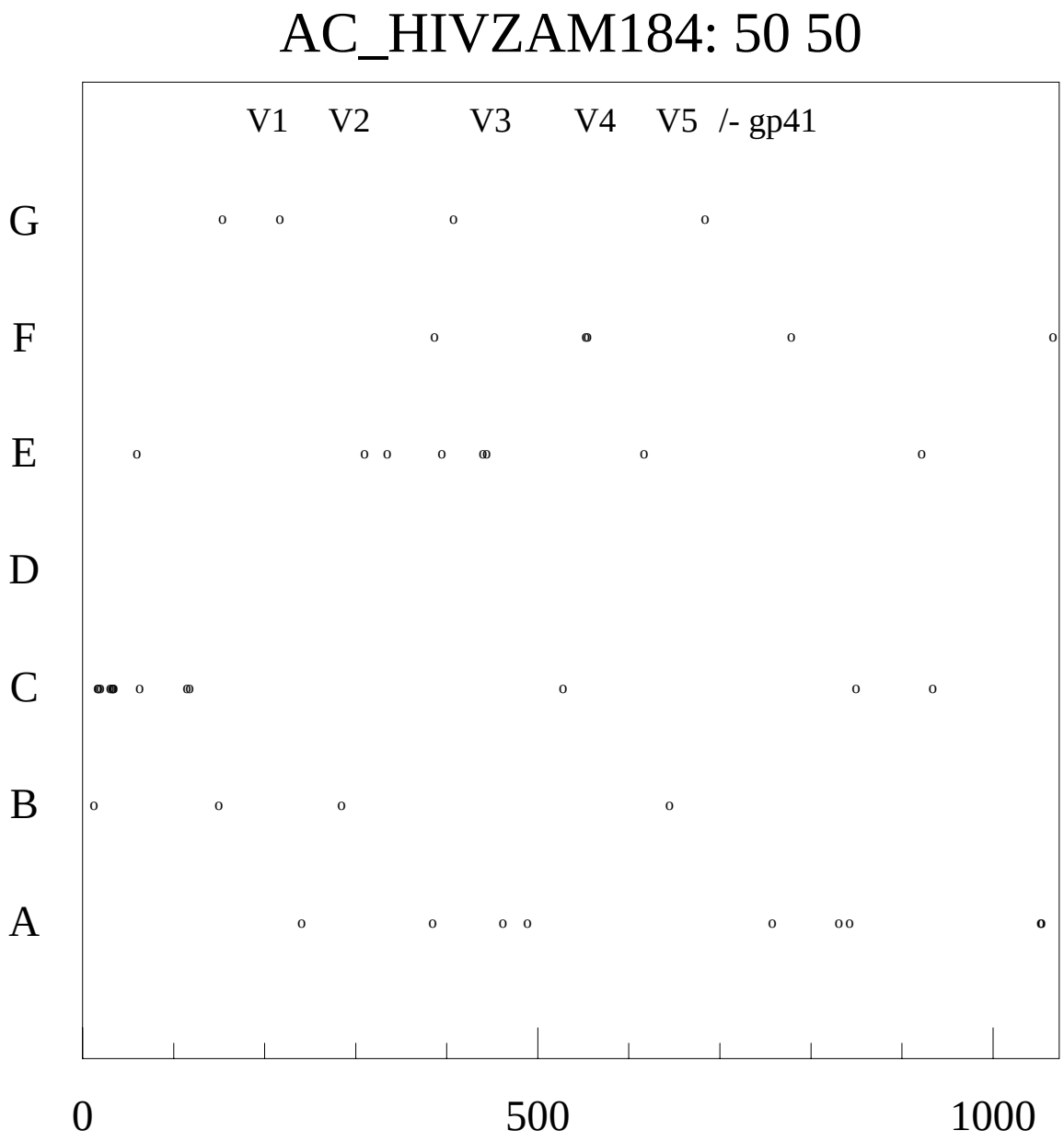


Fig. 1a. The putative hybrid ZAM184 compared to subtype signatures using constant thresholds for *p* and *q* of 50,50. ZAM184 appears to be hybrid.

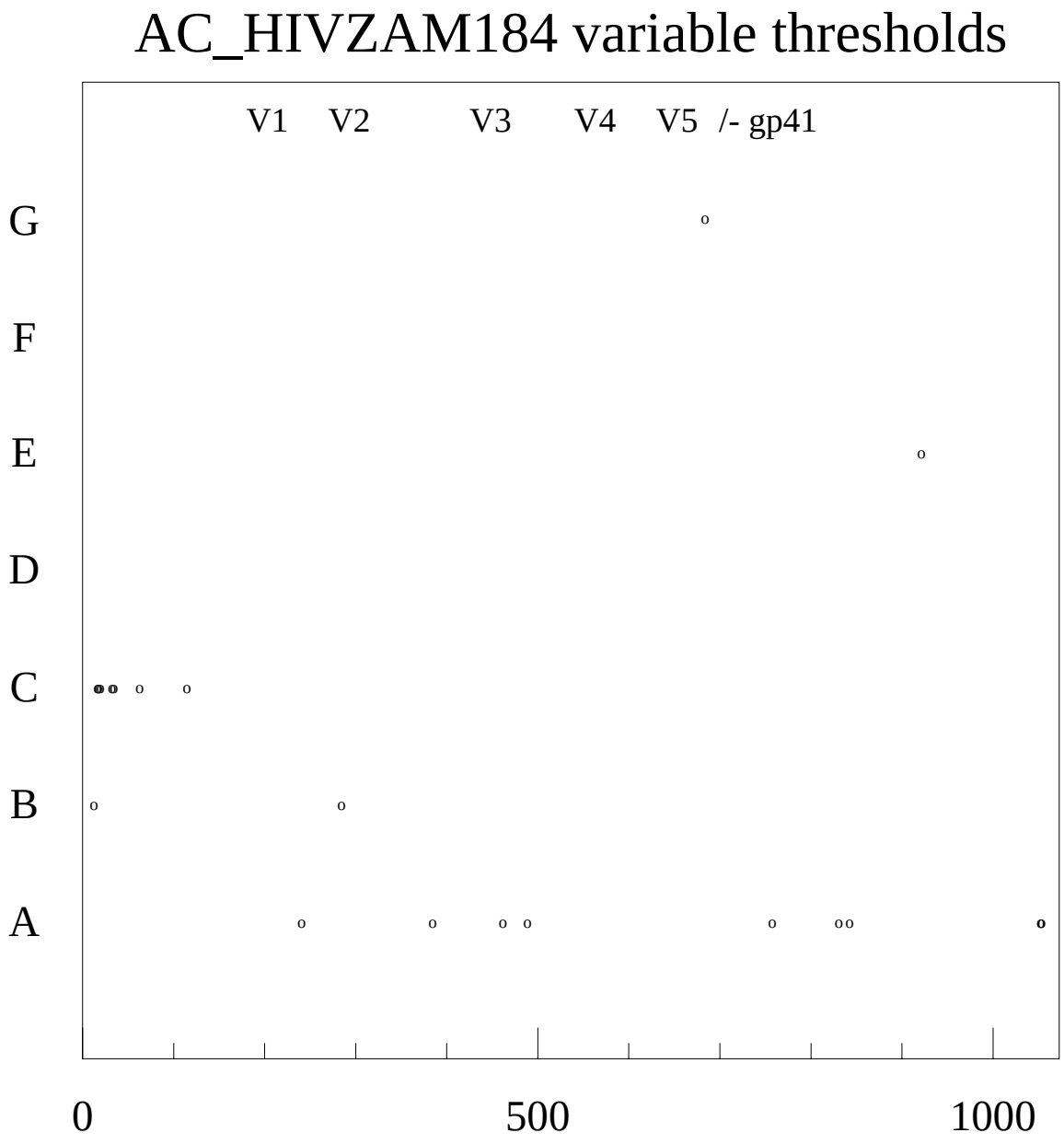
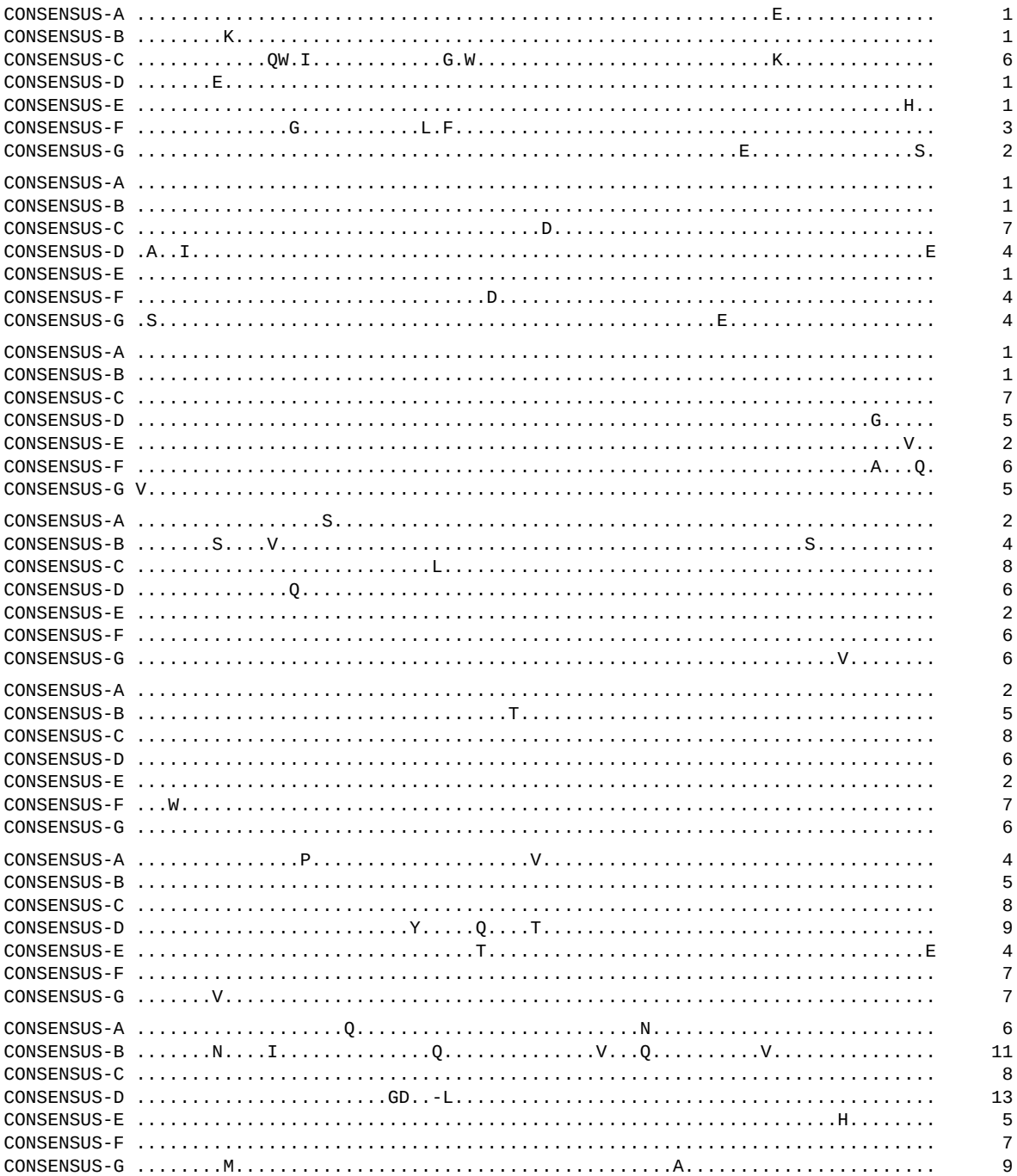


Fig. 1b. The putative hybrid ZAM184 compared to subtype signatures using variable thresholds that generate approximately equal number of signature sites for each of the seven subtypes. ZAM184 appears to be hybrid and homoplasy has been significantly reduced.

HIV Hybrids using VESPA

An alignment of unique VESPA-derived signature characters for HIV-1 sequence subtypes A-G is shown in Figure 2. The number of differentiating characters is relatively low (16–20), but we contend that the signal-to-noise ratio over these characters is high. In effect, we are saying that only a little over 100 of the 850 total amino acid residues are “informative” for purposes of analysis of potential recombinants. No other set of conditions for the available data set was found to be as satisfactory.

Fig. 2



```
CONSENSUS-A ..........................................................Q.........    7
CONSENSUS-B ..............................................................     11
CONSENSUS-C ..............................................................      8
CONSENSUS-D ..............................................................     13
CONSENSUS-E ..............................................................      5
CONSENSUS-F ..............................................................      7
CONSENSUS-G ...........................................................R..     10

CONSENSUS-A ..............Q...............................................      8
CONSENSUS-B ................Q.............................................     12
CONSENSUS-C ..............................................................      8
CONSENSUS-D ..............................................................     13
CONSENSUS-E ...........................................................I....      6
CONSENSUS-F ..............................................................      7
CONSENSUS-G ..............................................................     10

CONSENSUS-A ..............................................................      8
CONSENSUS-B ............................Q.................................     13
CONSENSUS-C ..............................................................      8
CONSENSUS-D ..............................................................     13
CONSENSUS-E .........Q.....................................I...............      8
CONSENSUS-F ......................Q.......................................      8
CONSENSUS-G ....................R.........G..................................V..     13

CONSENSUS-A ....................K.........................................      9
CONSENSUS-B ..............................................................     13
CONSENSUS-C ..................................I...........................      9
CONSENSUS-D ..........................................................RH......     15
CONSENSUS-E ..........................................KF..................T..     11
CONSENSUS-F ..............................................................      8
CONSENSUS-G ..............................................................     13

CONSENSUS-A ............S.......L...........I.............................     12
CONSENSUS-B ..................................T...........................     14
CONSENSUS-C ................................T.............................     10
CONSENSUS-D ..............................................................     15
CONSENSUS-E ..................................EI.T......D..................     15
CONSENSUS-F ............................S.E...............................     10
CONSENSUS-G ..............................................................     13

CONSENSUS-A ......................I.......................................     13
CONSENSUS-B ............V.....................................................G...     16
CONSENSUS-C ..........................................L...................     11
CONSENSUS-D ......................L.......................................     16
CONSENSUS-E ....................................P.........................     16
CONSENSUS-F ............................K...........S.....................     12
CONSENSUS-G ..............K...................................................     14

CONSENSUS-A ..............................................................R    14
CONSENSUS-B ..........................................................W.....     17
CONSENSUS-C ..........................................Q.....................V....L    14
CONSENSUS-D ...S..........................................................     17
CONSENSUS-E ..............................................................     16
CONSENSUS-F ......V.......................----.....................L......T.....     19
CONSENSUS-G ..................................NI..........................     16

CONSENSUS-A ......................IG......................                     16
CONSENSUS-B ......V.....................................                     18
CONSENSUS-C ...K.....................................A...                     16
CONSENSUS-D ...........................................                     17
CONSENSUS-E ...........................................                     16
CONSENSUS-F ......................A......................                     20
CONSENSUS-G ...........................................                     16
```

Fig. 2. Alignment of HIV-1 subtype signature patterns using variable thresholds.

A series of analyses of putative intragenic *env* mosaic sequences, UG266A, MAL, K124A, and DI2ACD, are shown in figures 3–6. Of these, a strong argument for mosaicism might be made for only K124A and DI2ACD (and ZAM184, as seen in Fig. 1b): either the signal is too low by the VESPA approach or the noise level is too high in the approaches that conclude these are hybrid molecules. Other parameter constraints, *i.e.* constant thresholds such as 50,50, gave stronger signals, however the homoplasy was high (Fig. 1a). Again, the focus herein is limited to detection and reduction of homoplasy. We envision that scoring strategies with the VESPA approach will require "runs" statistics. For a semi-quantitative evaluation of a newly-determined sequence, the alignment of differentiating characters in Figure 2 can be employed by sequencers.

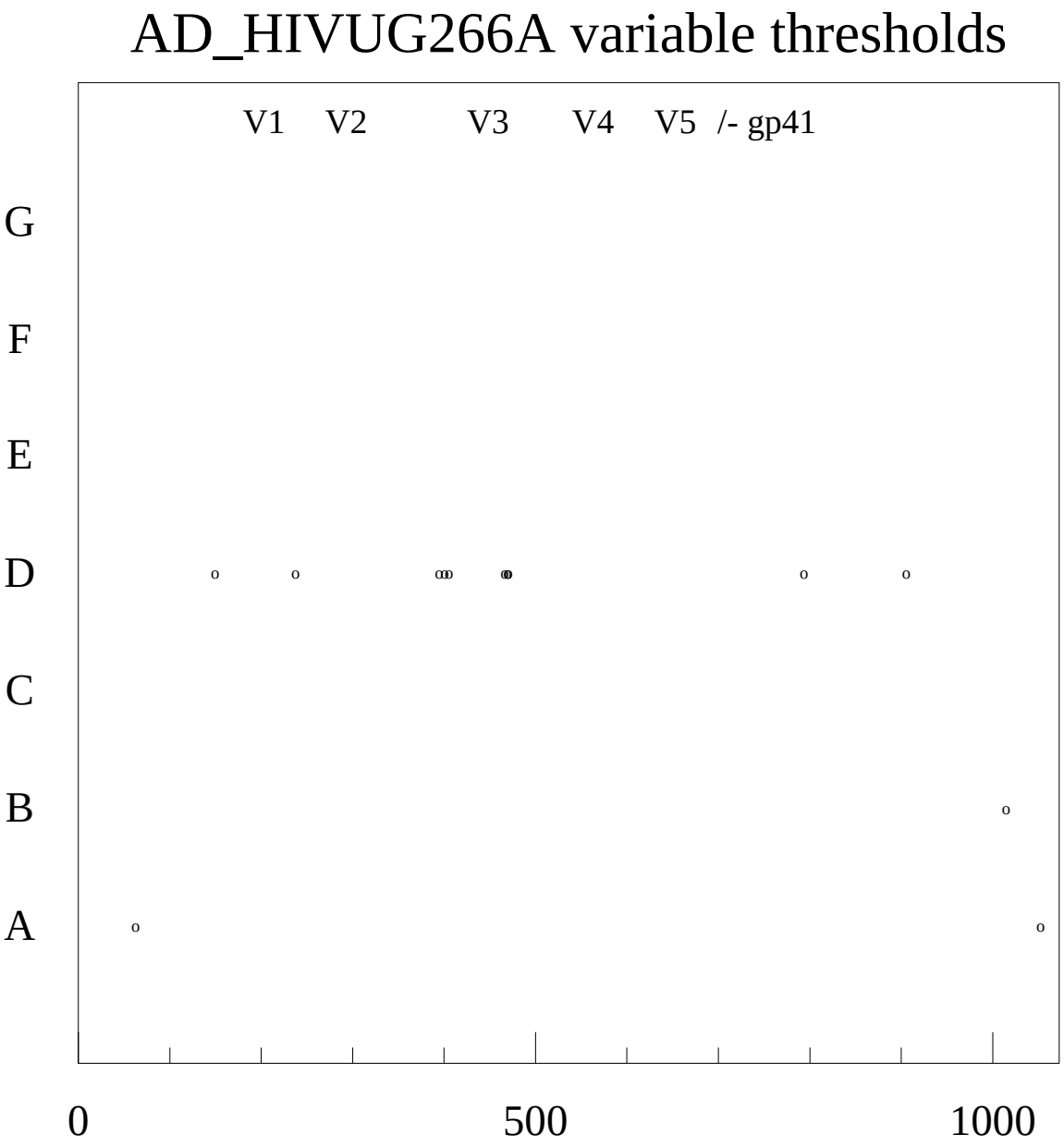


Fig. 3. UG266A, hypothesized to be an A-D intra-*env* recombinant, as analyzed by VESPA using variable thresholds.

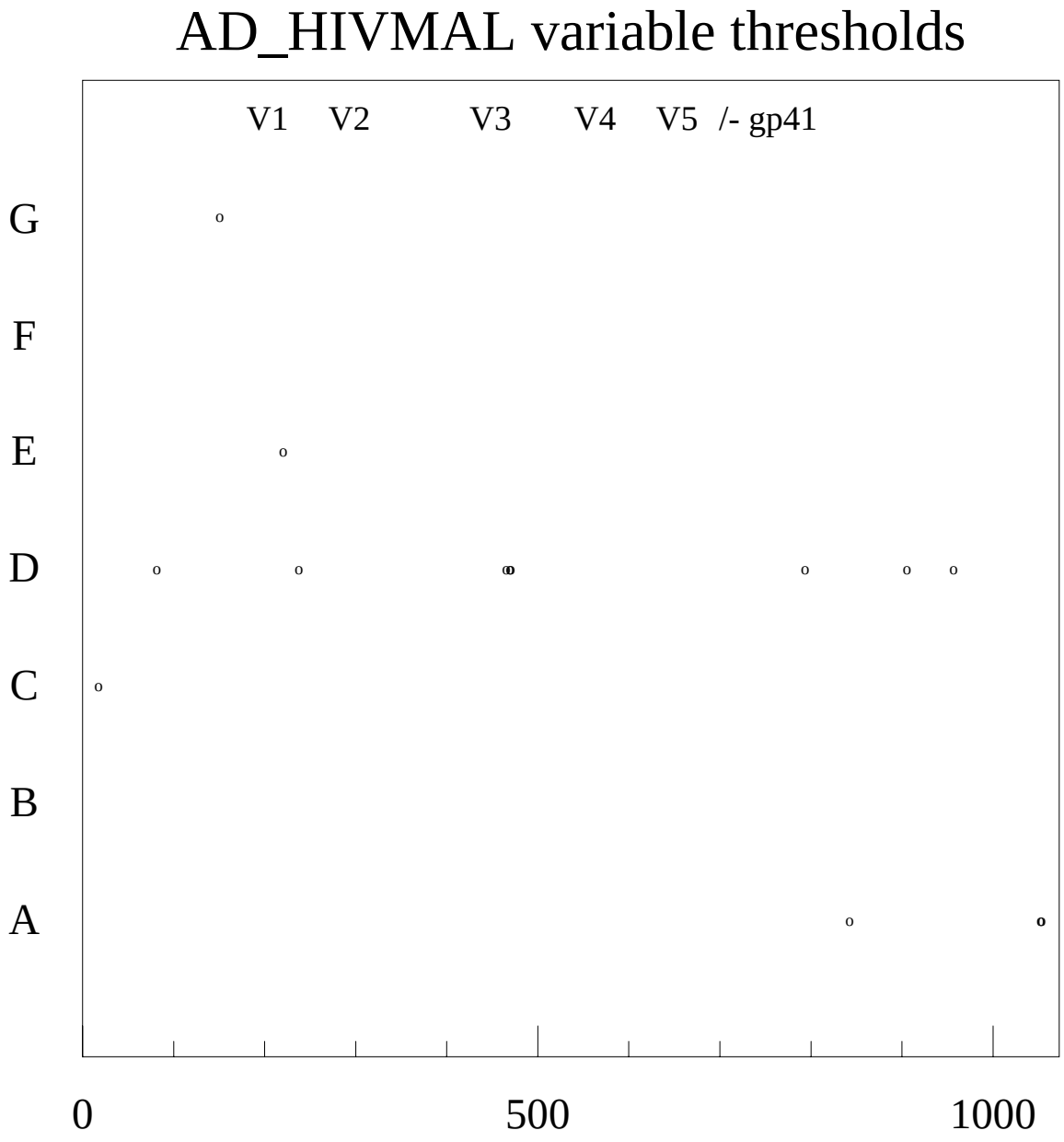


Fig. 4. MAL, hypothesized to be an A-D intra-*env* recombinant, as analyzed by VESPA using variable thresholds.

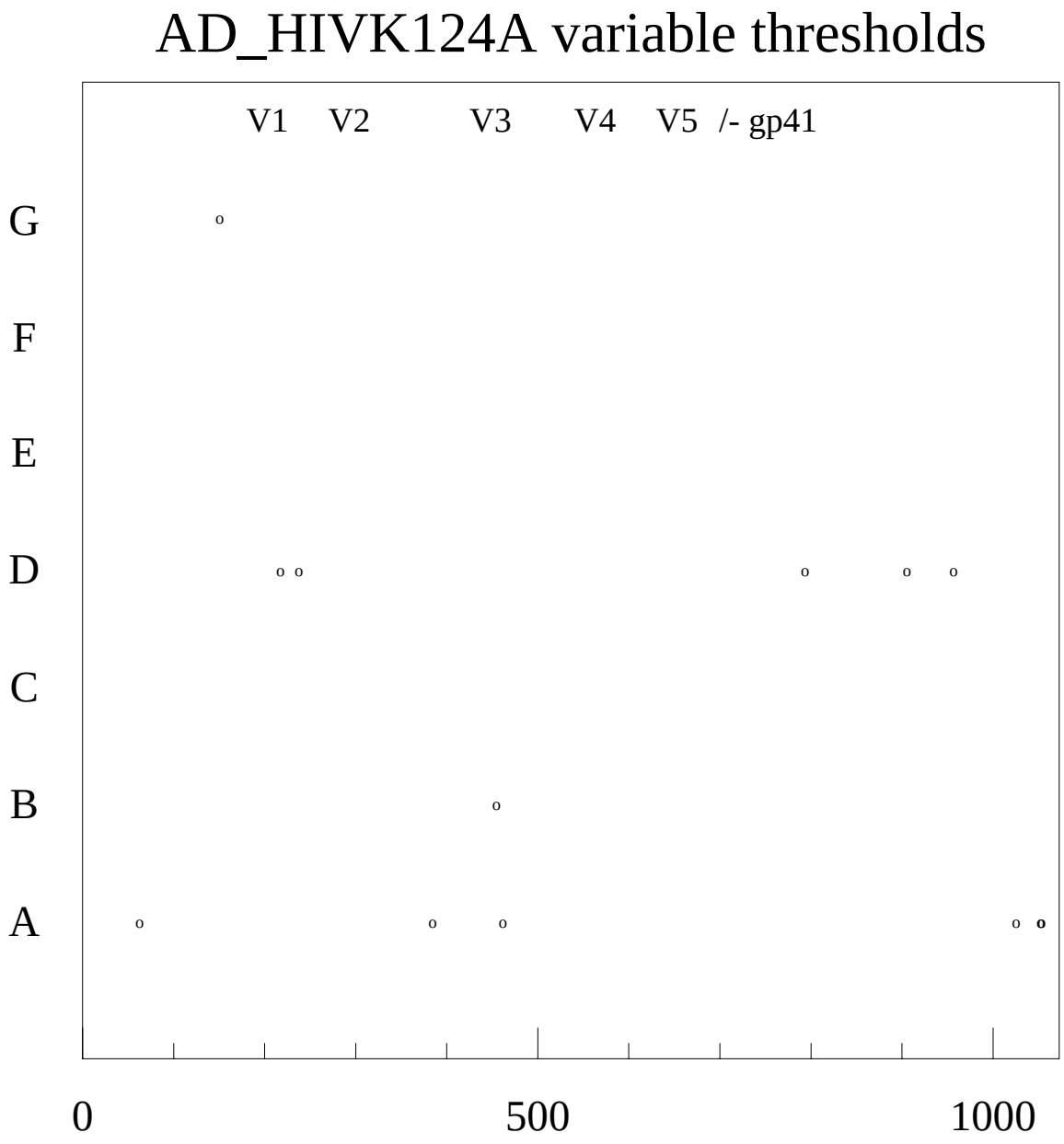


Fig. 5. K124A, hypothesized to be an A-D intra-*env* recombinant, as analyzed by VESPA using variable thresholds.

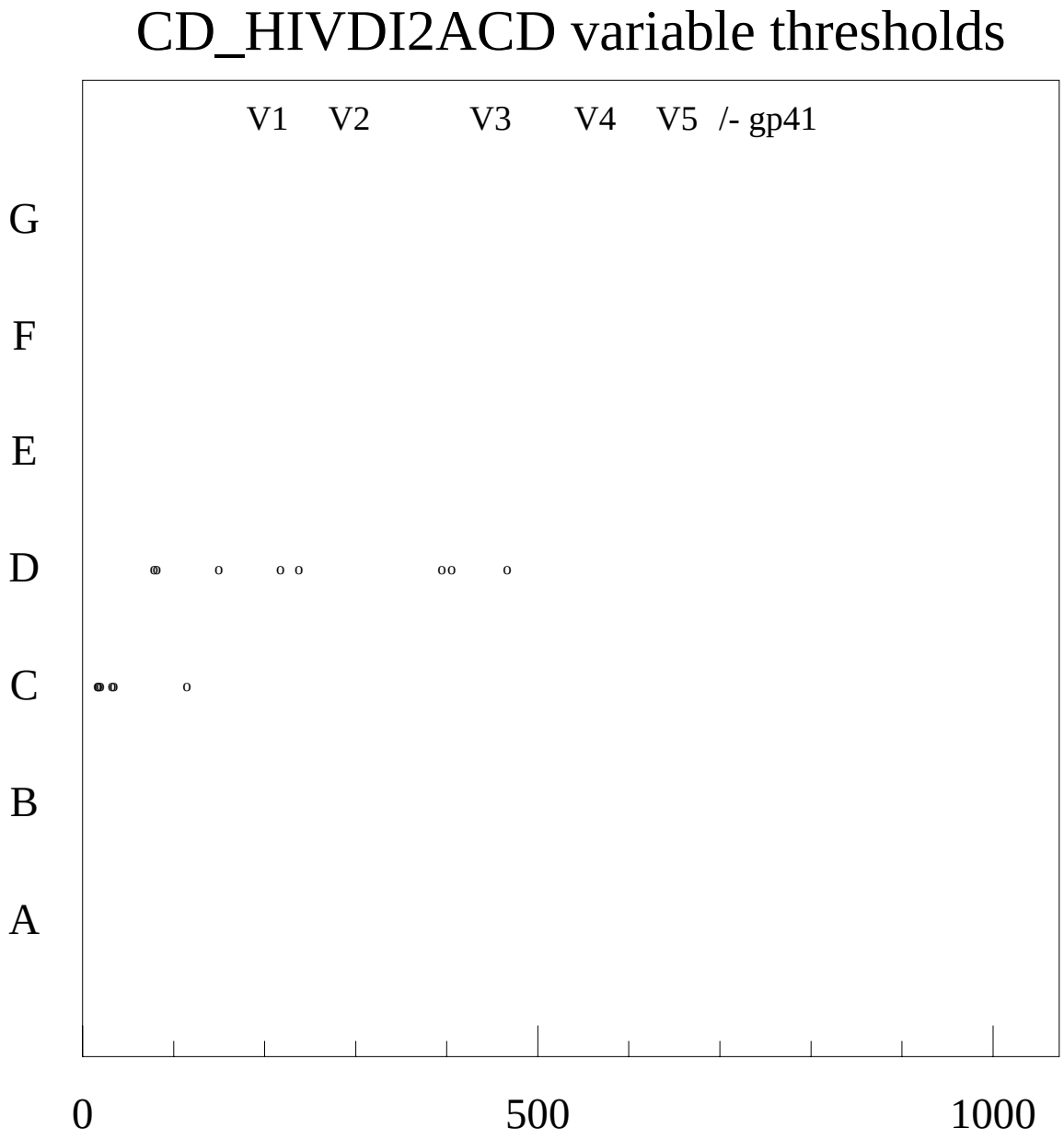


Fig. 6. DI2ACD, hypothesized to be an C-D intra-*env* recombinant, as analyzed by VESPA using variable thresholds. Note that DI2ACD has not been sequenced over gp41, which explains the absence of any signature matches over this region.

B. Intrasubtype Analysis

Fifty or so complete B subtype *env* sequences are available for hybrid analysis, however some of these are close siblings for which the phylogenetic similarities are too strong for separate inclusion. Restricting our attention to 40 B subtype sequences and using the constraints 75,20 for *r*, *q*, a table of the number of individual signature characters shared between different sequences was generated (Table 1). For this approach, we expect to see some strong sharing of signature characters between sequences due to phylogenetic linkage. Nevertheless, the differentiation seen in Table 1 is striking; the diagonal gives the number of signature residues for each of the sequences. The evaluation of potential recombinants from among the set of 40 sequences must begin by determining the fraction of sites shared with another sequence's diagonal. If a significant fraction of signature characters is shared, more than would obtain by "undirected convergence", this may indicate recombination; in most instances, it will indicate mere phylogenetic linkage. The **distribution** of shared sites must be statistically assessed in order to separate these alternatives, especially if a sequence is hybrid between something represented in the table and something not represented in the table.

P896	46	1	2	5	4	1	1	1	3	4	5	7	4	4	4	0	1	3	2	3	2	3	3	2	3	0	2	5	2	2	2	1	2	0	4	2	0	4	3	4
CDC42	1	40	2	3	1	0	2	1	0	2	2	4	1	5	1	2	3	1	2	2	0	1	0	0	1	1	3	4	3	1	2	1	1	1	3	0	2	1	2	2
HAN	2	2	41	2	2	5	4	0	0	1	1	3	4	1	2	0	1	3	2	2	2	1	3	2	5	0	5	2	1	5	2	4	3	2	3	3	3	1	2	3
SF33	5	3	2	33	4	3	1	1	0	3	2	2	1	3	1	0	0	1	2	2	0	1	1	0	1	0	1	1	0	0	3	2	2	0	1	1	2	5	2	8
3202A12	4	1	2	4	30	3	3	0	1	2	0	2	0	2	1	0	0	1	1	1	2	2	0	0	1	0	0	1	1	0	0	0	3	0	0	0	2	3	2	2
SF2	1	0	5	3	3	30	5	1	0	1	2	2	0	0	1	0	0	1	2	0	3	1	2	1	2	1	1	2	2	2	0	3	3	1	1	0	3	2	1	2
MANC	1	2	4	1	3	5	42	0	0	2	1	4	1	2	1	2	1	3	3	1	1	2	5	1	2	0	1	2	0	2	2	3	6	1	1	0	2	0	1	1
91US712.4	1	1	0	1	0	1	0	34	8	8	5	2	0	0	0	2	1	0	0	1	2	1	0	0	1	0	1	1	1	1	3	1	1	1	2	0	3	1	2	2
92US716.6	3	0	0	0	1	0	0	8	26	8	7	3	1	2	1	2	1	1	1	2	2	1	1	1	0	1	0	1	0	0	2	3	1	0	1	0	0	3	2	0
92US715.6	4	2	1	3	2	1	2	8	8	37	6	2	2	3	1	4	0	3	3	1	2	1	2	1	2	1	2	4	1	2	4	2	0	1	5	3	2	2	4	3
MN	5	2	1	2	0	2	1	5	7	6	34	3	4	1	2	5	1	0	1	3	2	3	2	2	1	1	0	4	2	2	4	2	2	1	2	1	1	1	3	3
168A	7	4	3	2	2	2	4	2	3	2	3	36	1	3	1	3	2	2	2	3	4	0	2	1	2	2	0	2	3	3	2	0	4	1	1	2	5	3	5	3
BRVA	4	1	4	1	0	0	1	0	1	2	4	1	31	3	0	3	2	0	1	1	2	1	4	2	1	0	1	3	3	2	2	1	1	2	2	0	0	2	0	0
WEAU160	4	5	1	3	2	0	2	0	2	3	1	3	3	30	1	2	1	2	0	1	2	1	0	0	4	3	3	5	2	3	4	0	1	0	4	1	2	2	2	2
WMJ22	4	1	2	1	1	1	1	0	1	1	2	1	0	1	27	0	0	1	2	3	1	0	0	0	0	0	0	1	1	0	1	1	1	0	0	2	0	0	2	4	2
BAL1	0	2	0	0	0	0	2	2	2	4	5	3	3	2	0	22	3	1	1	1	0	0	1	1	1	1	0	4	2	1	1	1	0	2	3	1	1	0	1	1
LAI	1	3	1	0	0	0	1	1	1	0	1	2	2	1	0	3	23	0	1	1	0	1	0	0	0	0	0	3	1	0	0	0	0	2	0	0	2	0	1	1
D31	3	1	3	1	1	1	3	0	1	3	0	2	0	2	1	1	0	27	2	1	0	1	2	0	4	0	1	1	1	0	0	2	1	1	2	0	0	1	1	0
ALA1	2	2	2	2	1	2	3	0	1	3	1	2	1	0	2	1	1	2	24	2	1	1	0	0	1	2	3	2	0	1	2	2	4	1	1	1	2	2	2	1
JH32	3	2	2	2	1	0	1	1	2	1	3	3	1	1	3	1	1	1	2	29	1	2	1	1	1	1	3	1	0	3	1	2	3	0	1	0	2	1	4	1
YU2	2	0	2	0	2	3	1	2	2	2	2	4	2	2	1	0	0	0	1	1	14	2	1	1	3	2	2	0	1	2	3	0	1	0	1	0	1	0	2	0
ENVVA	3	1	1	1	2	1	2	1	1	1	3	0	1	1	0	0	1	1	1	2	2	16	2	1	0	1	0	0	0	0	0	1	3	0	0	0	1	0	1	0
JRCSF	3	0	3	1	0	2	5	0	1	2	2	2	4	0	0	1	0	2	0	1	1	2	19	5	1	0	1	1	2	0	1	3	1	1	0	0	1	0	1	0
JRFL	2	0	2	0	0	1	1	0	1	1	2	1	2	0	0	1	0	0	0	0	1	1	1	5	13	0	0	1	1	1	0	0	2	2	1	0	0	0	0	0	0
ADA	3	1	5	1	1	2	2	1	0	2	1	2	1	4	0	1	0	4	1	1	3	0	1	0	19	2	1	1	3	5	2	4	1	0	2	1	3	0	1	0
TH1412	0	1	0	0	0	1	0	0	1	1	1	2	0	3	0	1	0	0	0	2	1	2	1	0	0	2	23	3	2	0	2	1	0	1	0	1	1	3	2	3	2
SF162	2	3	5	1	0	1	1	1	0	2	0	0	1	3	1	0	0	1	3	3	2	0	1	1	1	3	20	2	1	6	4	2	2	1	1	0	1	1	2	0
US3	5	4	2	1	1	2	2	1	1	4	4	2	3	5	1	4	3	1	2	1	0	0	1	1	1	2	2	27	3	5	4	1	1	1	3	1	1	4	1	1
NY5CG	2	3	1	0	1	2	0	1	0	1	2	3	3	2	0	2	1	1	0	0	1	0	2	1	3	0	1	3	23	4	3	2	1	1	1	1	1	1	1	0
CAM1	2	1	5	0	0	2	2	1	0	2	2	3	2	3	1	1	0	0	1	3	2	0	0	0	5	2	6	5	4	29	6	1	1	1	2	0	2	1	1	3
SIMI84	2	2	2	3	0	0	2	3	2	4	4	2	2	4	1	1	0	0	2	1	3	0	1	0	2	1	4	4	3	6	37	1	1	0	2	2	3	1	1	3
BCSG3C	1	1	4	2	0	3	3	1	3	2	2	0	1	0	1	1	0	2	2	2	0	1	3	2	4	0	2	1	2	1	1	31	3	2	3	0	1	0	1	3
OYI	2	1	3	2	3	3	6	1	1	0	2	4	1	1	0	0	0	1	4	3	1	3	1	2	1	1	2	1	1	1	1	3	36	3	0	3	3	3	2	6
92HT593.1	0	1	2	0	0	1	1	1	0	1	1	1	2	0	0	2	2	1	1	0	0	0	1	1	0	0	1	1	1	1	0	2	3	29	1	3	4	1	1	2
91HT652.11	4	3	3	1	0	1	1	2	1	5	2	1	2	4	2	3	0	2	1	1	1	0	0	0	2	1	1	3	1	2	2	3	0	1	34	5	1	2	6	0
91HT651.11	2	0	3	1	0	0	0	0	0	3	1	2	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	1	1	0	2	0	3	3	5	33	3	0	1	3
92HT596.4	0	2	3	2	2	3	2	3	0	2	1	5	0	2	0	1	2	0	2	2	1	1	1	0	3	3	1	1	1	2	3	1	3	4	1	3	31	2	3	6
RF	4	1	1	5	3	2	0	1	3	2	1	3	2	2	2	0	0	1	2	1	0	0	0	0	0	2	1	4	1	1	1	0	3	1	2	0	2	31	3	3
US4	3	2	2	2	2	1	1	2	2	4	3	5	0	2	4	1	1	1	2	4	2	1	1	0	1	3	2	1	1	1	1	1	2	1	6	1	3	3	43	4
92HT599.24	4	2	3	8	2	2	1	2	0	3	3	3	0	2	2	1	1	0	1	1	0	0	0	0	0	2	0	1	0	3	3	3	6	2	0	3	6	3	4	63

Table 1. A summary of individual signature characters for 40 B subtype Env sequences using the constraints 75,20. The number of sites for each character pattern is shown on the diagonal; the numbers off the diagonal indicate the number of shared characters for each signature. An alignment of the characters is presented on the database Web site, *http://hiv-web.lanl.gov*.

In the absence at this time of any definitive B-B hybrid, we have constructed a chimeric sequence of two of the 40 sequences, which involves two crossover sites. A comparison of this chimera with the 40 reference signatures was made for *r*, *q* equal to 75,20 (Fig. 7a) and 75,10 (Fig. 7b). The shift in the signal-to-noise ratio is dramatic: while both analyses support the mosaic nature of the chimeric sequence, the noise level (homoplasy) is high for the less stringent setting.

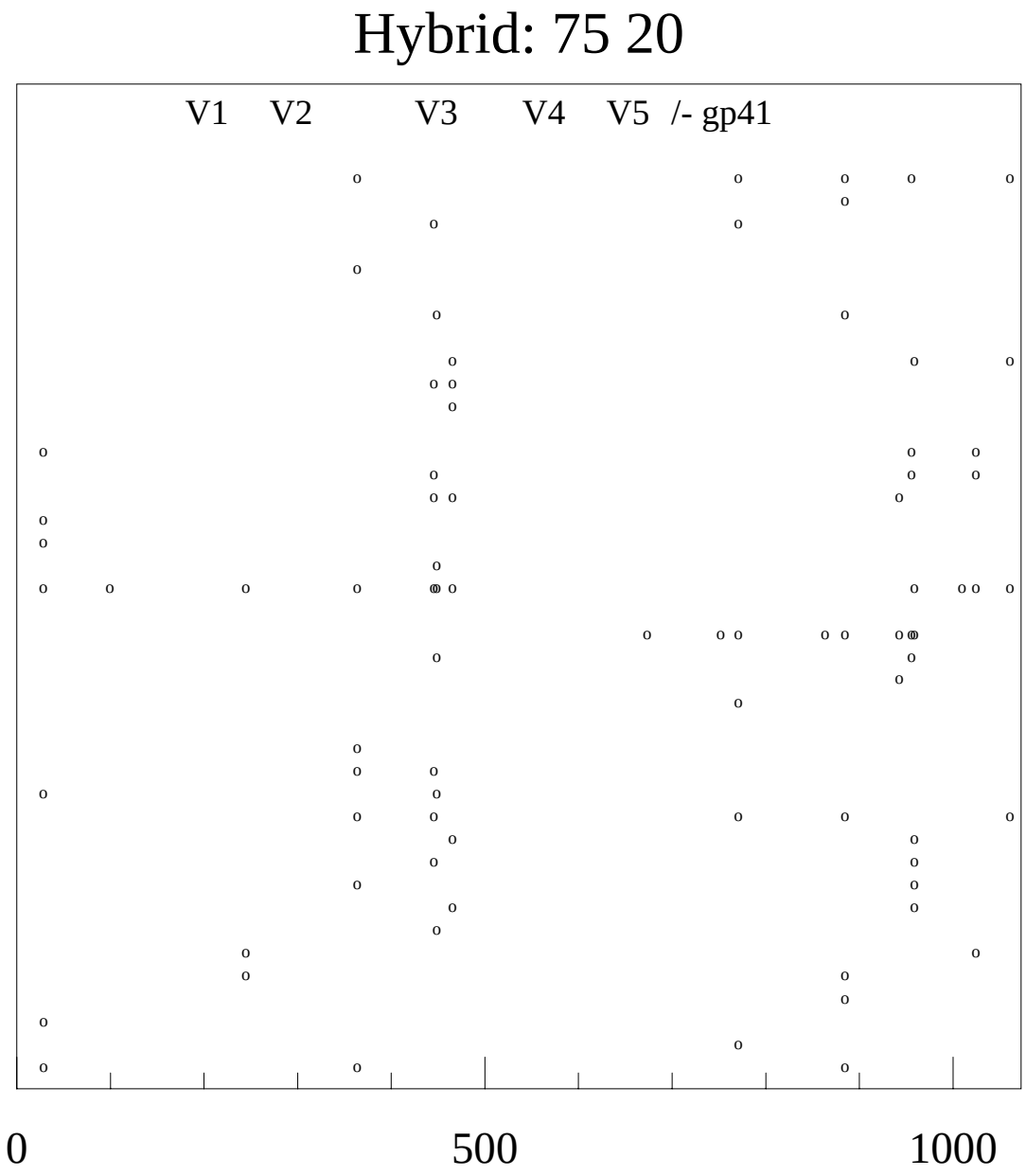


Fig. 7. A computer-constructed B-B intra-*env* recombinant, as analyzed by VESPA using *r*, *q* settings of 75,20 (a) and 75,10 (b). Matches between the chimeric sequence and individual reference sequences are evident in each row. Two rows of matches at the center of the plot reveal the true sources of the chimera and the two crossover events.

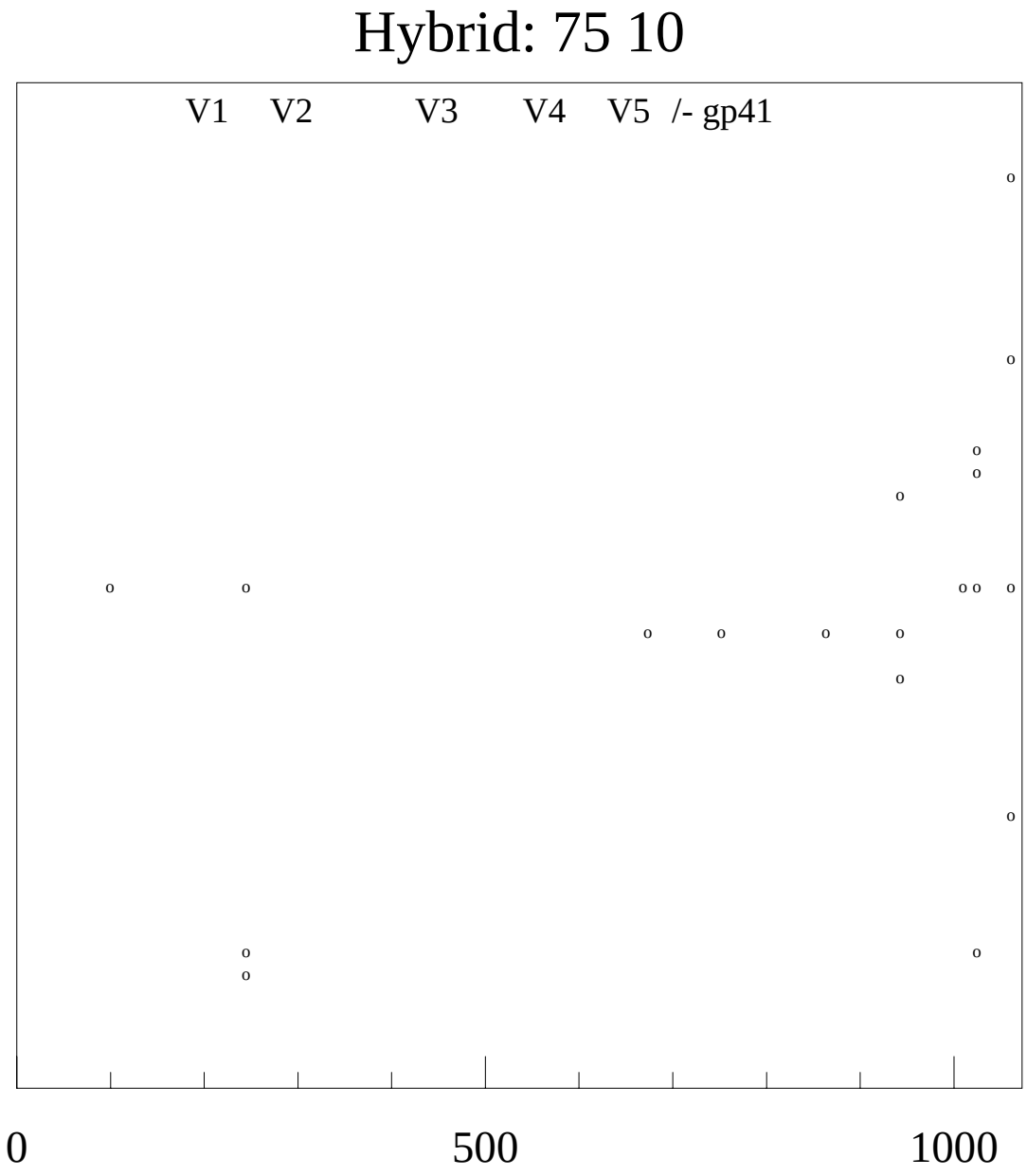


Fig. 7. A computer-constructed B-B intra-*env* recombinant, as analyzed by VESPA using *r*, *q* settings of 75,20 (a) and 75,10 (b).

CONCLUDING REMARKS

A continuation of this investigation will involve the analysis of Gag sequences as well as Env sequences and the application of various scoring methods. Our preliminary results suggest that homoplasy is a pervasive problem with hybrid analysis, and that methods blind to this fact may overestimate the extent of HIV-1 recombination. Some thought must be given by all methods to mosaicism that arises from **lack of divergence** and from **convergence** rather than from recombination as such.

References

Farmer, A.D. and Myers, G. (1995) In *Human Papillomaviruses 1995*, edited by Myers, G., Bernard, H.-U., Delius, H., et al.), Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, 1995, pp. III-139–III-146.

Korber,B. and Myers,G. (1992) Signature Pattern Analysis: a Method for Assessing Viral Sequence Relatedness. *AIDS Research and Human Retroviruses* **8**:1549–1560.

Li,W.-H. and Graur,D. *Fundamentals of Molecular Evolution*, Sinauer Associates, Sunderland, 1991, pp. 111–113.

Ou, C.-Y., Ciesielski,C., Myers,G., et al. (1992) Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**:1165–1171.

Stewart, C.-B. (1993) The powers and pitfalls of parsimony. *Nature* **361**:603–607.

Wills, C. (1995) Topiary Pruning of the HIV and SIV Phylogenetic Tree. *AIDS Research and Human Retroviruses* **11**:1417–1419.

APPENDIX I

To ensure uniqueness of signature characters for each of the subtypes, a necessary and sufficient condition is $p \geq q$,
where p = minimum frequency for signature character in query set
q = upper limit on frequency for signature character in background set
Proof:

In order for a character to be shared by two subtypes (X and Y), it must be a signature character for those subtypes considered against all other subtypes (so that it will be included in the consensus), including each other (X as query against Y as background, and vice versa). Thus, using X as the query set and Y as its background, a character (a) will be a signature character for X against Y if and only if

$$X(a) \geq p \text{ and } Y(a) < q.$$

$X(a)$ denotes the frequency of character a in set X.
Using Y as the query against X as a background, (a) will be a signature character for Y against X if and only if

$$Y(a) \geq p \text{ and } X(a) < q.$$

Thus, in order for the character to be included in the signature pattern for X and for Y, we must have

$$(p \leq X(a) < q) \text{ and } (p \leq Y(a) < q)$$

clearly this is impossible if and only if $p \geq q$
If using different threshold values for the different subtypes, the character a will be a signature character for both X and Y if and only if

$$X(a) \geq p1 \text{ and } Y(a) < q1;$$
$$Y(a) \geq p2 \text{ and } X(a) < q2$$

implying

$$(p1 \leq X(a) < q2) \text{ and } (p2 \leq Y(a) < q1)$$

which will be impossible if and only if

$$(p1 \geq q2) \text{ or } (p2 \geq q1).$$